







# Heterogeneous graph contrastive learning with gradient balance for drug repositioning

Hai Cui <sup>1,†</sup>, Meiyu Duan <sup>1,†</sup>, Haijia Bi <sup>2</sup>, Xiaobo Li <sup>1</sup>, Xiaodi Hou <sup>1</sup>, Yijia Zhang <sup>1,\*</sup>

<sup>1</sup>Information Science and Technology College, Dalian Maritime University, No.1 Linghai Road, Dalian 116026, Liaoning, China

<sup>2</sup>College of Computer Science and Technology, Jilin University, No.2699 Qianjin Street, Changchun 130012, Jilin, China

\*Corresponding author. Information Science and Technology College, Dalian Maritime University, No.1 Linghai Road, Dalian 116026, Liaoning, China.

E-mail: zhangyijia@dlmu.edu.cn

<sup>†</sup>Hai Cui and Meiyu Duan contributed equally to this work.

## Abstract

Drug repositioning, which involves identifying new therapeutic indications for approved drugs, is pivotal in accelerating drug discovery. Recently, to mitigate the effect of label sparsity on inferring potential drug–disease associations (DDAs), graph contrastive learning (GCL) has emerged as a promising paradigm to supplement high-quality self-supervised signals through designing auxiliary tasks, then transfer shareable knowledge to main task, i.e. DDA prediction. However, existing approaches still encounter two limitations. The first is how to generate augmented views for fully capturing higher-order interaction semantics. The second is the optimization imbalance issue between auxiliary and main tasks. In this paper, we propose a novel heterogeneous Graph Contrastive learning method with Gradient Balance for DDA prediction, namely GCGB. To handle the first challenge, a fusion view is introduced to integrate both semantic views (drug and disease similarity networks) and interaction view (heterogeneous biomedical network). Next, inter-view contrastive learning auxiliary tasks are designed to contrast the fusion view with semantic and interaction views, respectively. For the second challenge, we adaptively adjust the gradient of GCL auxiliary tasks from the perspective of gradient direction and magnitude for better guiding parameter update toward main task. Extensive experiments conducted on three benchmarks under 10-fold cross-validation demonstrate the model effectiveness.

**Keywords:** drug repositioning; heterogeneous information network; multi-task learning; graph contrastive learning; gradient-based optimization

## Introduction

To solve the dilemma of high investment and low success rate in new drug development, computational-based drug repositioning (DR), which targets at learning from multi-source heterogeneous biological data and identifying new therapeutic indications for approved or late-stage clinical trial drugs, has gradually emerged as a complementary and promising solution.

Since the drug–disease association (DDA) network can be naturally formed as a bipartite graph structure, growing efforts have been devoted to exploiting the advantages of graph neural networks (GNNs) for effectively integrating multiple biological relationships. Existing GNN-based solutions are generally divided into three mainstream branches from the perspective of biological network types.

- (1) *Similarity network-based methods*. Under the assumption that similar drugs are more inclined to treat similar diseases and vice versa [1], this line of researches first calculates the similarity scores between each pair of drugs and diseases through diverse measurement modalities, then the obtained similarity features are applied to predict underlying DDAs [2, 3].

- (2) *Association network-based methods*. This line focuses on propagating and aggregating information from heterogeneous association networks, which comprise multiple types of biological entities and various interaction relationships, e.g. drug–protein and protein–disease etc., to capture potential associations between drugs and diseases [4–6].
- (3) *Dual networks fusion-based methods*. Above two kinds of networks are not mutually exclusive, to benefit from both of their advantages, emerging approaches design dedicated model architecture to integrate similarity network with association network, thereby generating more comprehensive representations for drugs and disease [7–9].

Despite their widespread applications, the performance of GNN-based DDA approaches is susceptible to the scale of labeled training samples. Unfortunately, due to the labor-intensive and time-consuming process of wet experiments, the number of validated DDAs is insufficient compared with the holistic interaction space, which is also known as the issue of *label sparsity* [10, 11]. Specifically, we assume that the numbers of drugs, diseases, and drug–disease positive samples are  $m$ ,  $n$ , and  $k$ , respectively, thus there is a guarantee that  $k \ll m \times n$  holds.

Received: September 18, 2024. Revised: November 2, 2024. Accepted: November 29, 2024

© The Author(s) 2024. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<https://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited.

For commercial re-use, please contact [journals.permissions@oup.com](mailto:journals.permissions@oup.com)

Under such circumstances, plagued by label sparsity, prior GNN-based solutions fall short of fully capturing intricate semantic correlations between drugs and diseases.

To mitigate the effect of sparse supervised signals, graph contrastive learning (GCL), as the representative of self-supervised learning techniques, has been introduced for DDA prediction [12, 13]. The general processing flow involves firstly performing graph augmentation to generate different views of original graph. Afterwards, contrastive objectives are defined to maximize and minimize the consistency of positive as well as negative node pairs between the original and augmented views, respectively. At last, the GCL tasks served as auxiliary tasks are jointly optimized with the main task (i.e. DDA prediction) based on multi-task learning paradigm [14].

Although GCL shines in combating against the issue of label sparsity, existing approaches devised for DDA prediction still suffer from two limitations. (i) Random perturbation strategies, including node and edge dropout, noisy representations etc., are frequently utilized for graph augmentation [5, 12, 15]. However, as demonstrated in literature [13], such strategies lead to performance degradation. We argue that the augmented views built by random perturbation fail to capture higher-order interaction semantics (i.e. drug-drug and disease-disease relationships) implicit in the original DDA network. (ii) The imbalance phenomenon of gradient direction and magnitude between GCL auxiliary tasks and main task is detrimental to the prediction accuracy of target task [16]. More concretely, let  $\mathcal{L}_{\text{GCL}}^{(i)}$  and  $\mathcal{L}_{\text{DDA}}$  denote the loss of  $i$ -th GCL auxiliary task and main task respectively,  $\theta$  refers to the shared parameters, which is optimized by jointly multi-task training.  $\mathbf{G}_{\text{GCL}}^{(i)}$  and  $\mathbf{G}_{\text{DDA}}$  denote the corresponding gradients w.r.t.  $\theta$ , i.e.  $\mathbf{G}_{\text{GCL}}^{(i)} = \nabla_{\theta} \mathcal{L}_{\text{GCL}}^{(i)}$ ,  $\mathbf{G}_{\text{DDA}} = \nabla_{\theta} \mathcal{L}_{\text{DDA}}$ . Intuitively, larger gradient magnitude ( $L_2$  norm,  $\|\cdot\|$ ) dominates the overall optimization trend. Thus, if  $\|\mathbf{G}_{\text{GCL}}^{(i)}\| \gg \|\mathbf{G}_{\text{DDA}}\|$ , the optimizer inclines to update shared parameters toward the  $i$ -th GCL auxiliary task rather than main task, resulting in serious issue of optimization imbalance (please refer to Section Adaptive Gradient Balance for details). However, till now, this issue is seldom considered in the field of DR. To sum up, how to design meaningful GCL auxiliary tasks and solve the optimization imbalance issue are two intractable challenges.

To deal with the first challenge, distinct from random perturbation, we expect to construct separate drug and disease similarity networks, which are deemed as *semantic views*, while we consider the DDA network as *interaction view*. Instead of directly generating contrastive sample pairs between the semantic and interaction views, a *fusion view* is introduced to integrate both higher-order relationships (drug-drug and disease-disease) and interaction information (drug-disease). Afterwards, we decide to contrast the fusion view with semantic and interaction views, respectively. Such a contrastive paradigm guarantees the mutual information between paired nodes across views is maximized.

As for the second challenge, in order to prevent GCL auxiliary tasks from dominating the optimization process, a simple and straightforward solution is to introduce weight hyperparameters for each auxiliary task, i.e.  $\mathbf{G}_{\text{GCL}}^{(i)'} = \lambda_i \nabla_{\theta} \mathcal{L}_{\text{GCL}}^{(i)}$ , where  $\lambda_i$  is a hyperparameter. However, tuning the weights for multiple GCL tasks by grid or random search is extremely time-consuming. More importantly, since the gradient magnitudes are dynamically changing during the training process, such fixed task hyperparameters trap in local optimum. Inspired by prior works [17, 18], in this paper, we explore to flexibly adapt the gradient of GCL auxiliary tasks from the perspective of gradient direction and magnitude for better transferring knowledge to main task.

To this end, we propose a novel heterogeneous Graph Contrastive learning method with Gradient Balance for identifying potential DDAs (abbreviated as GCGB). GCGB consists of four crucial components, i.e. node representation learning, DDA predictor, inter-view contrastive learning (CL), and adaptive gradient balance.

Specifically, drug and disease similarity networks are constructed in advance. Moreover, to enrich graph connectivity, the original DDA network integrated with two protein-related bipartite graphs (i.e. drug-protein and disease-protein) composes a heterogeneous interaction network. Subsequently, the constructed similarity networks and interaction network, which are treated as semantic views and interaction view respectively, are taken as the input of the first module. This module firstly utilizes dual graph transformer (GT) networks [19, 20] to perform message propagation and aggregation on these views, then at each GT layer, multi-head self-attention mechanism [21] is employed to fuse node representations from both semantic and interaction views. Next, the second module introduces two additional composition operators to calculate the probability that a given drug is effective in treating a specific disease. After that, the third module designs inter-view CL auxiliary tasks to alleviate label sparsity. More concretely, let the final fused drug and disease embeddings produced by mean-pooling across all GT layers be the fusion view. We contrast the fusion view with semantic and interaction views, respectively. The positive samples are the identical nodes within different views, while the negative ones are the disparate nodes within different views. Finally, the fourth module adaptively alters the gradient of above auxiliary tasks by simultaneously considering gradient direction and magnitude. Briefly, if the gradient magnitude of  $i$ -th GCL auxiliary task is larger than the counterpart of main task, i.e.  $\|\mathbf{G}_{\text{GCL}}^{(i)}\| > \|\mathbf{G}_{\text{DDA}}\|$ , we detect if the directions of these two gradients are conflicting, and further rectify  $\mathbf{G}_{\text{GCL}}^{(i)}$  by measuring the magnitude proximity with  $\mathbf{G}_{\text{DDA}}$ .

In a nutshell, the main contributions of this paper can be summarized as follows:

- A novel heterogeneous GCL method with gradient balance, namely GCGB, is proposed for inferring potential DDAs. To the best of our knowledge, it is the first time that optimization imbalance phenomenon between GCL auxiliary tasks and main task is considered in DR.
- We design effective inter-view CL auxiliary tasks through contrasting the fusion view with semantic and interaction views respectively, thereby maximizing the mutual information between paired nodes across views.
- To prevent auxiliary tasks from dominating the optimization process, we adaptively alter the gradient of GCL auxiliary tasks from the perspective of gradient direction and magnitude for better guiding parameter update toward main task.
- Extensive experiments demonstrate that GCGB outperforms the competitive baselines on three commonly-used benchmarks under 10-fold cross-validation. Furthermore, we also conduct detailed case studies to predict candidate drugs for different diseases, rendering it a practical and trustworthy tool for DR.

## Materials and preliminaries

### Datasets

We verify the effectiveness of our proposed method on three commonly-used benchmarks, i.e. B-dataset [22], C-dataset [23], and F-dataset [24]. Please refer to the supplementary material

Table 1. Statistic of three benchmark datasets

Dataset	Node number			Association number			Sparsity
	Drug	Disease	Protein	Drug-Disease	Drug-Protein	Disease-Protein	
B-dataset	269	598	1021	18416	3110	5898	11.45%
C-dataset	663	409	993	2532	3672	10691	0.93%
F-dataset	592	313	2741	1933	3152	47470	1.04%

for details of these benchmarks. Table 1 shows the statistic of above datasets. Specifically, let  $m$ ,  $n$ , and  $k$  denote the number of drugs, diseases, and DDAs, respectively, the *sparsity* ratio is calculated as  $\frac{k}{m \times n}$ . We observe that all datasets encounter varying degrees of sparsity, C-dataset and F-dataset are much sparser than B-dataset. Hence, how to effectively mitigate the issue of label sparsity is essential for DDA prediction.

### Biological networks construction

Herein, we will successively introduce the DDA network, heterogeneous interaction network, and homogeneous similarity network.

**DDA network.** Let  $\mathcal{V}^{\text{DR}}$  and  $\mathcal{V}^{\text{DI}}$  denote the set of drugs and diseases, respectively. The DDA network is deemed as an undirected bipartite graph  $\mathcal{G}_{\text{DDA}} = \{\mathcal{V}^{\text{DR}}, \mathcal{V}^{\text{DI}}, \mathcal{E}_{\text{DDA}}\}$ , where  $\mathcal{E}_{\text{DDA}} \subseteq \mathcal{V}^{\text{DR}} \times \mathcal{V}^{\text{DI}}$  denotes the associations that have been experimentally validated between drugs and diseases. An adjacency matrix w.r.t.  $\mathcal{G}_{\text{DDA}}$  is represented as  $\mathbf{A}_{\text{DDA}} \in \{0, 1\}^{|\mathcal{V}^{\text{DR}}| \times |\mathcal{V}^{\text{DI}}|}$ , where  $\mathbf{A}_{\text{DDA}}^{uv} = 1$  if drug  $u$  can treat disease  $v$ , otherwise  $\mathbf{A}_{\text{DDA}}^{uv} = 0$ .

**Heterogeneous interaction network.** It can be treated as an undirected heterogeneous information network, which consists of three node types, i.e. drug (DR for short), disease (DI), and protein (PR), and three edge types, i.e. drug-disease (DDA), drug-protein (DRP), and disease-protein (DIP). Formally, let  $\mathcal{V}^{\text{PR}}$  denote the set of proteins. The heterogeneous interaction network is defined as  $\mathcal{G}_{\text{HIN}} = \{\mathcal{V}, \mathcal{E}_{\text{HIN}}, \mathcal{T}^{\text{V}}, \mathcal{T}^{\text{E}}\}$ , where  $\mathcal{V} = \mathcal{V}^{\text{DR}} \cup \mathcal{V}^{\text{DI}} \cup \mathcal{V}^{\text{PR}}$  and  $\mathcal{E}_{\text{HIN}} = \mathcal{E}_{\text{DDA}} \cup \mathcal{E}_{\text{DRP}} \cup \mathcal{E}_{\text{DIP}}$  refer to the set of nodes and edges, respectively. Moreover, two type mapping functions, i.e.,  $\phi : \mathcal{V} \rightarrow \mathcal{T}^{\text{V}}$  and  $\psi : \mathcal{E}_{\text{HIN}} \rightarrow \mathcal{T}^{\text{E}}$  assign the corresponding type to each node and edge, where  $\mathcal{T}^{\text{V}} = \{\text{DR}, \text{DI}, \text{PR}\}$ ,  $\mathcal{T}^{\text{E}} = \{\text{DDA}, \text{DRP}, \text{DIP}\}$ .  $\mathcal{G}_{\text{HIN}}$  could be represented by a series of adjacency matrices  $\{\mathbf{A}_e : e \in \mathcal{T}^{\text{E}}\}$ . Taking the edge type DRP  $\in \mathcal{T}^{\text{E}}$  as an example,  $\mathbf{A}_{\text{DRP}} \in \{0, 1\}^{|\mathcal{V}^{\text{DR}}| \times |\mathcal{V}^{\text{PR}}|}$  is an adjacency matrix where nonzero values indicate the existence of interactions between drugs and proteins.

**Homogeneous similarity network.** For drug similarity network, it is denoted as  $\mathcal{G}_{\text{DR}}$ , and the corresponding adjacency matrix is  $\mathbf{A}_{\text{DR}} \in \{0, 1\}^{|\mathcal{V}^{\text{DR}}| \times |\mathcal{V}^{\text{DR}}|}$ , where  $\mathbf{A}_{\text{DR}}^{uj} = 1$  if drug  $j$  is the top-K nearest neighbor of drug  $u$ , otherwise  $\mathbf{A}_{\text{DR}}^{uj} = 0$ . In practice, following prior works [7, 25], we comprehensively calculate both the drug fingerprint similarity and Gaussian interaction profile (GIP) kernel similarity for each drug pair. Analogously, the disease similarity network is denoted as  $\mathcal{G}_{\text{DI}}$ , the adjacency matrix w.r.t.  $\mathcal{G}_{\text{DI}}$  is represented as  $\mathbf{A}_{\text{DI}} \in \{0, 1\}^{|\mathcal{V}^{\text{DI}}| \times |\mathcal{V}^{\text{DI}}|}$ , where  $\mathbf{A}_{\text{DI}}^{vj} = 1$  if disease  $j$  is the top-K nearest neighbor of disease  $v$ , otherwise  $\mathbf{A}_{\text{DI}}^{vj} = 0$ . We also simultaneously measure disease phenotype similarity as well as GIP kernel similarity for each disease pair.

### Methodology

The overall architecture of GCGB is shown in Fig. 1. In the following subsections, we will elaborate on each component.

### Node representation learning

The constructed similarity networks ( $\mathcal{G}_{\text{DR}}$ ,  $\mathcal{G}_{\text{DI}}$ ) and interaction network ( $\mathcal{G}_{\text{HIN}}$ ) are deemed as semantic views and interaction view, respectively. We separate the overall procedure of node representation learning into three sub-steps, i.e. semantic view feature extraction, interaction view feature extraction, and layer-wise feature fusion.

#### Semantic view feature extraction

Since semantic views are homogeneous information networks in essence, we employ the basic GT network [19] to obtain drug and disease representations from  $\mathcal{G}_{\text{DR}}$  and  $\mathcal{G}_{\text{DI}}$ , respectively.

Taking  $\mathcal{G}_{\text{DR}}$  as an example, let  $\mathbf{h}_{\text{DR},u}^{(l)} \in \mathbb{R}^d$  denote the  $d$ -dimensional embedding of drug  $u$  at  $l$ -th layer on  $\mathcal{G}_{\text{DR}}$ . We exploit multi-head attention mechanism to estimate the importance of each neighbor node and aggregate the neighborhood message by attention weights. Concretely, as for the drug node  $u$  and its neighbor node  $j \in \mathcal{N}_{\mathcal{G}_{\text{DR}}}(u)$ , drug  $u$  is transformed into a query vector, i.e.  $\mathbf{Q}_h^{(l)} = \mathbf{W}_{Q,h}^{(l)} \mathbf{h}_{\text{DR},u}^{(l)}$ , while neighbor node  $j$  is mapped into a key vector, i.e.  $\mathbf{K}_h^{(l)} = \mathbf{W}_{K,h}^{(l)} \mathbf{h}_{\text{DR},j}^{(l)}$  and a value vector, i.e.  $\mathbf{V}_h^{(l)} = \mathbf{W}_{V,h}^{(l)} \mathbf{h}_{\text{DR},j}^{(l)}$ , where  $h$  refers to  $h$ -th head and the total head number is  $H$ ,  $\mathbf{W}_{Q,h}^{(l)}$ ,  $\mathbf{W}_{K,h}^{(l)}$ , and  $\mathbf{W}_{V,h}^{(l)}$  are learnable parameter matrices. Afterwards, the scaled dot-product attention of  $h$ -th head is calculated to capture the correlation between query and key vectors:

$$\text{ATT}_{\text{GT}}^h(j, u) = \text{Softmax}_{\mathbf{v} \in \mathcal{N}_{\mathcal{G}_{\text{DR}}}(u)} \left( \frac{\mathbf{K}_h^{(l)\top} \mathbf{Q}_h^{(l)}}{\sqrt{d/H}} \right) \quad (1)$$

where  $\mathcal{N}_{\mathcal{G}_{\text{DR}}}(u)$  denotes all the neighbor nodes of drug  $u$  on  $\mathcal{G}_{\text{DR}}$ . Next, the message from node  $j$  to  $u$  is formulated as the concatenation of weighted multi-head value vectors:

$$\text{MSG}_{\text{GT}}(j, u) = \parallel_{h=1}^H \text{ATT}_{\text{GT}}^h(j, u) \mathbf{V}_h^{(l)} \quad (2)$$

where  $\parallel$  refers to the concatenation operation. Finally, the  $(l+1)$ -th layer embedding of drug  $u$  is generated by utilizing message passing mechanism to aggregate information from all neighbor nodes and adding the  $\mathbf{h}_{\text{DR},u}^{(l)}$  as a residual term:

$$\mathbf{h}_{\text{DR},u}^{(l+1)} = \sigma \left( \sum_{\mathbf{v} \in \mathcal{N}_{\mathcal{G}_{\text{DR}}}(u)} \mathbf{W}_{\text{GT}}^{(l)} \text{MSG}_{\text{GT}}(j, u) \right) + \mathbf{h}_{\text{DR},u}^{(l)} \quad (3)$$

where  $\sigma(\cdot)$  denotes a non-linear activation function,  $\mathbf{W}_{\text{GT}}^{(l)}$  is an optimizable parameter. After the calculation of  $L$  layers, we preserve the node representations of all layers for each drug  $\{\mathbf{h}_{\text{DR},u}^{(l)} | l = [1, \dots, L]\}_{u=0}^{|\mathcal{V}^{\text{DR}}|}$ .

Analogously, as for the disease similarity network  $\mathcal{G}_{\text{DI}}$ , we apply another basic GT to propagate and aggregate information from neighbor nodes. The corresponding node representations of all layers for each disease are denoted as  $\{\mathbf{h}_{\text{DI},v}^{(l)} | l = [1, \dots, L]\}_{v=0}^{|\mathcal{V}^{\text{DI}}|}$ .

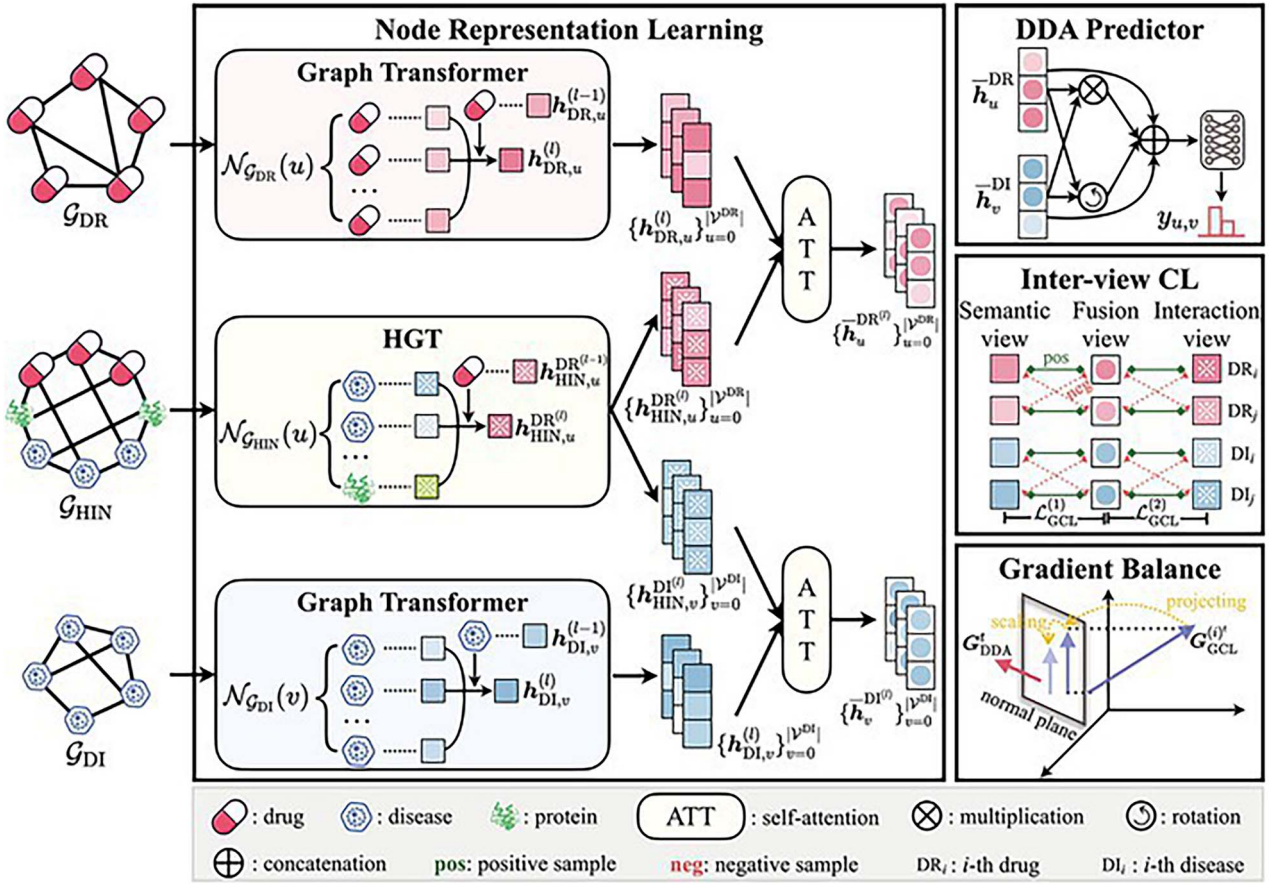


Figure 1. The overall architecture of GCGB, which consists of four crucial components, i.e. node representation learning, DDA predictor, inter-view CL, and adaptive gradient balance. The primary innovations lie in the design of inter-view CL and adaptive gradient balance modules. Specifically, as for inter-view CL, a fusion view is introduced to integrate both higher-order relationships (drug-drug and disease-disease) and interaction information (drug-disease). Afterwards, we contrast the fusion view with semantic and interaction views respectively, thereby maximizing the mutual information between paired nodes across views. As for adaptive gradient balance, we dynamically adjust the gradient of GCL auxiliary tasks through reducing the proportion of conflicting gradient directions and measuring the proximity of gradient magnitudes at each training epoch, effectively achieving parameter update toward main task.

### Interaction view feature extraction

The interaction view is modeled as a heterogeneous information network. However, the basic GT assumes that different types of nodes/edges share the identical feature space, making it infeasible to capture heterogeneous properties. In order to maintain graph heterogeneity, we resort to heterogeneous graph transformer (HGT) [20] to project each node/edge type with a specific transformation matrix.

Concretely, let  $\mathbf{h}_{\text{HIN},i}^{(l)} \in \mathbb{R}^d$  denote the  $d$ -dimensional embedding of node  $i$  at  $l$ -th layer on  $\mathcal{G}_{\text{HIN}}$ . Given an edge  $e = (j, i) \in \mathcal{E}_{\text{HIN}}$  linked from source node  $j$  to target node  $i$ , the corresponding edge and node types are  $\psi(e)$ ,  $\phi(j)$ , and  $\phi(i)$ , we first project the representations of target node  $i$  and source node  $j$  into multi-head query, key, and value vectors, respectively, i.e.  $\mathbf{Q}_h^{(l)'} = \mathbf{W}_{Q,h}^{(\phi(i))} \mathbf{h}_{\text{HIN},i}^{(l)}$ ,  $\mathbf{K}_h^{(l)'} = \mathbf{W}_{K,h}^{(\phi(j))} \mathbf{h}_{\text{HIN},j}^{(l)}$  and  $\mathbf{V}_h^{(l)'} = \mathbf{W}_{V,h}^{(\phi(j))} \mathbf{h}_{\text{HIN},j}^{(l)}$ , where  $h$  denotes  $h$ -th head and the total head number is also set to  $H$ ,  $\mathbf{W}_{Q,h}^{(\phi(i))}$ ,  $\mathbf{W}_{K,h}^{(\phi(j))}$ , and  $\mathbf{W}_{V,h}^{(\phi(j))}$  are node type-specific parameter matrices. Afterwards, we calculate the  $h$ -th head attention for each edge  $e$  as follows:

$$\text{ATT}_{\text{HGT}}^h(j, e, i) = \text{Softmax}_{\psi(j) \in \mathcal{N}_{\mathcal{G}_{\text{HIN}}}(i)} \left( \frac{\mathbf{K}_h^{(l)'} \mathbf{W}_{\psi(e)}^{\text{A}^{(l)}} \mathbf{Q}_h^{(l)'}}{\sqrt{d/H}} \right) \quad (4)$$

where  $\mathcal{N}_{\mathcal{G}_{\text{HIN}}}(i)$  refers to all the neighbor nodes of target  $i$  on  $\mathcal{G}_{\text{HIN}}$ .  $\mathbf{W}_{\psi(e)}^{\text{A}^{(l)}}$  is an edge type-specific parameter. Next, the edge type is

incorporated into message passing process, thus, the message from node  $j$  to  $i$  through edge  $e$  is formulated as:

$$\text{MSG}_{\text{HGT}}(j, e, i) = \bigoplus_{h=1}^H \text{ATT}_{\text{HGT}}^h(j, e, i) \mathbf{W}_{\psi(e)}^{\text{M}^{(l)}} \mathbf{V}_h^{(l)'} \quad (5)$$

where  $\mathbf{W}_{\psi(e)}^{\text{M}^{(l)}}$  is also an edge type-specific parameter. Finally, we calculate the  $(l+1)$ -th layer embedding of node  $i$  by aggregating information from all neighbor nodes, followed by the non-linear activation and residual connection:

$$\mathbf{h}_{\text{HIN},i}^{(l+1)} = \sigma \left( \sum_{\psi(j) \in \mathcal{N}_{\mathcal{G}_{\text{HIN}}}(i)} \mathbf{W}_{\phi(i)}^{(l)} \text{MSG}_{\text{HGT}}(j, e, i) \right) + \mathbf{h}_{\text{HIN},i}^{(l)} \quad (6)$$

where  $\mathbf{W}_{\phi(i)}^{(l)}$  is a node type-specific parameter. After stacking the HGT blocks for  $L$  layers, the highly contextualized representation is produced for each node. For ease of description, the obtained vector representations of all layers for drug and disease nodes are denoted as  $\{\mathbf{h}_{\text{HIN},u}^{\text{DR}(l)} | l = [1, \dots, L]\}_{u=0}^{|\mathcal{V}^{\text{DR}}|}$  and  $\{\mathbf{h}_{\text{HIN},v}^{\text{DI}(l)} | l = [1, \dots, L]\}_{v=0}^{|\mathcal{V}^{\text{DI}}|}$ , respectively.

### Layer-wise feature fusion

After obtaining the drug and disease representations from both semantic and interaction views, a fusion view is introduced to



integrate above two views at each layer through multi-head self-attention mechanism.

Specifically, taking drug  $u$  as an example, at  $l$ -th layer, the corresponding node embeddings learned from semantic and interaction views are  $\mathbf{h}_{\text{DR},u}^{(l)}$  and  $\mathbf{h}_{\text{HIN},u}^{\text{DR}(l)}$ , respectively. Afterwards, multi-head self-attention mechanism is utilized to fuse these two views, the fused representation of drug  $u$  at  $l$ -th layer is denoted as  $\bar{\mathbf{h}}_u^{\text{DR}(l)}$ . Analogously, as for disease  $v$ , we denote the fused node representation at  $l$ -th layer as  $\bar{\mathbf{h}}_v^{\text{DI}(l)}$ . Finally, we exploit mean-pooling to combine the fused representations across all layers, thereby generating the final fused drug and disease embeddings:

$$\bar{\mathbf{h}}_u^{\text{DR}} = \frac{1}{L} \sum_{l=1}^L \bar{\mathbf{h}}_u^{\text{DR}(l)}, \bar{\mathbf{h}}_v^{\text{DI}} = \frac{1}{L} \sum_{l=1}^L \bar{\mathbf{h}}_v^{\text{DI}(l)} \quad (7)$$

### DDA predictor

Since the task of DDA prediction is normally regarded as a binary classification problem, DDA predictor targets at calculating the probability score that a given drug is able to treat a specific disease.

Specifically, given a drug-disease pair  $(u, v) \in \mathcal{V}^{\text{DR}} \times \mathcal{V}^{\text{DI}}$ , instead of directly concatenating the fused drug embedding  $\bar{\mathbf{h}}_u^{\text{DR}}$  with the disease embedding  $\bar{\mathbf{h}}_v^{\text{DI}}$  and then feeding it into the multi-layer perceptron (MLP), inspired by the idea of composition operators [26], we introduce the multiplication ( $\xi_{\text{Mul}}$ ) [27] and rotation ( $\xi_{\text{Rot}}$ ) [28] operators to further incorporate drug and disease embeddings.  $\xi_{\text{Mul}}$  performs the element-wise vector product, while  $\xi_{\text{Rot}}$  projects embedding vectors to complex space and utilizes the rotation of complex domain to describe relations.

Thus, the concatenation of  $\bar{\mathbf{h}}_u^{\text{DR}}, \bar{\mathbf{h}}_v^{\text{DI}}, \xi_{\text{Mul}}(\bar{\mathbf{h}}_u^{\text{DR}}, \bar{\mathbf{h}}_v^{\text{DI}})$  (for short) and  $\xi_{\text{Rot}}(\bar{\mathbf{h}}_u^{\text{DR}}, \bar{\mathbf{h}}_v^{\text{DI}})$  (for short) is fed into MLP to estimate the therapeutic probability:

$$y_{u,v} = \text{sigmoid}(\text{MLP}(\bar{\mathbf{h}}_u^{\text{DR}}, \bar{\mathbf{h}}_v^{\text{DI}}, \xi_{\text{Mul}}(\bar{\mathbf{h}}_u^{\text{DR}}, \bar{\mathbf{h}}_v^{\text{DI}}), \xi_{\text{Rot}}(\bar{\mathbf{h}}_u^{\text{DR}}, \bar{\mathbf{h}}_v^{\text{DI}}))) \quad (8)$$

Finally, binary cross-entropy is adopted as the loss function for DDA prediction:

$$\mathcal{L}_{\text{DDA}} = - \sum_{(u,v) \in S} \hat{y}_{u,v} \log y_{u,v} + (1 - \hat{y}_{u,v}) \log(1 - y_{u,v}) \quad (9)$$

where  $S$  represents the positive and negative training samples,  $\hat{y}_{u,v}$  refers to the ground-truth label.

### Inter-view contrastive learning

This module contrasts the fusion view with semantic and interaction views respectively, and ensures that the mutual information between paired nodes across views is maximized.

Specifically, we also apply mean-pooling to integrate the node embeddings across all layers, the corresponding representations of node  $i$  on  $\mathcal{G}_{\text{DR}}, \mathcal{G}_{\text{DI}}$ , and  $\mathcal{G}_{\text{HIN}}$  are denoted as  $\mathbf{h}_{\text{DR},i} = \frac{1}{L} \sum_{l=1}^L \mathbf{h}_{\text{DR},i}^{(l)}$ ,  $\mathbf{h}_{\text{DI},i} = \frac{1}{L} \sum_{l=1}^L \mathbf{h}_{\text{DI},i}^{(l)}$ , and  $\mathbf{h}_{\text{HIN},i}^{\phi(l)} = \frac{1}{L} \sum_{l=1}^L \mathbf{h}_{\text{HIN},i}^{\phi(l)}$ , respectively. Afterwards, we contrast the fusion view (Section Layer-wise Feature Fusion) with semantic views (i.e.  $\mathcal{G}_{\text{DR}}$  and  $\mathcal{G}_{\text{DI}}$ , Section Semantic View Feature Extraction). The identical nodes within different views are treated as positive samples, while the negative samples are distinct nodes within different views. It is worth noting that we remove the top- $K$  nearest neighbors of node  $i$  as negative instances, since false negatives (i.e. highly similar nodes) discard

### Algorithm 1 Overall procedure of parameter update

---

```

1: Input: Shared parameters  $\theta$ , learning rate  $\alpha$ , moving average coefficient  $\beta$ , relax factor  $\omega$ , total auxiliary tasks  $N$ , total epochs  $T$ .
2: Output: Updated shared parameters  $\theta^T$ .
3: Initialize  $m_{\text{DDA}}^0 = m_{\text{GCL}}^0 = 0$ 
4: for  $t = 1$  to  $T$  do
5:    $\mathbf{G}_{\text{DDA}}^t \leftarrow \nabla_{\theta} \mathcal{L}_{\text{DDA}}^t$ 
6:   Calculate  $m_{\text{DDA}}^t$  by Eq. 13
7:   for  $i = 1$  to  $N$  do
8:      $\mathbf{G}_{\text{GCL}}^{(i)t} \leftarrow \nabla_{\theta} \mathcal{L}_{\text{GCL}}^{(i)t}$ 
9:     Calculate  $m_{\text{GCL}}^{(i)t}$  by Eq. 13
10:    if  $m_{\text{GCL}}^{(i)t} > m_{\text{DDA}}^t$  then
11:      if  $\cos(\mathbf{G}_{\text{DDA}}^t, \mathbf{G}_{\text{GCL}}^{(i)t}) < 0$  then
12:        Alter gradient direction of  $\mathbf{G}_{\text{GCL}}^{(i)t}$  by Eq. 14
13:      end if
14:      Alter gradient magnitude of  $\mathbf{G}_{\text{GCL}}^{(i)t}$  by Eq. 14
15:    end if
16:  end for
17:   $\mathbf{G}_{\text{total}}^t \leftarrow \mathbf{G}_{\text{DDA}}^t + \sum_{i=1}^N \mathbf{G}_{\text{GCL}}^{(i)t}$ 
18:   $\theta^{t+1} \leftarrow \theta^t - \alpha \times \mathbf{G}_{\text{total}}^t$ 
19: end for
20: Return:  $\theta^T$ 

```

---

the true semantic information. Accordingly, the inter-view contrastive loss between fusion view and semantic views can be formulated as the following InfoNCE loss [29]:

$$\begin{aligned} \mathcal{L}_{\text{GCL}}^{(1)} = & - \sum_{u \in \mathcal{V}^{\text{DR}}} \log \frac{\exp((\bar{\mathbf{h}}_u^{\text{DR}})^T \mathbf{h}_{\text{DR},u})/\tau}{\sum_{j \in \{\mathcal{V}^{\text{DR}} - \mathcal{N}_{\mathcal{G}_{\text{DR}}}(u)\}} (\exp((\bar{\mathbf{h}}_u^{\text{DR}})^T \mathbf{h}_{\text{DR},j})/\tau)} \\ & - \sum_{v \in \mathcal{V}^{\text{DI}}} \log \frac{\exp((\bar{\mathbf{h}}_v^{\text{DI}})^T \mathbf{h}_{\text{DI},v})/\tau}{\sum_{j \in \{\mathcal{V}^{\text{DI}} - \mathcal{N}_{\mathcal{G}_{\text{DI}}}(v)\}} (\exp((\bar{\mathbf{h}}_v^{\text{DI}})^T \mathbf{h}_{\text{DI},j})/\tau)} \end{aligned} \quad (10)$$

where  $\tau$  is a temperature coefficient. Analogously, the inter-view contrastive loss between fusion view and interaction view (i.e.  $\mathcal{G}_{\text{HIN}}$ , Section Interaction View Feature Extraction) is calculated as follows:

$$\begin{aligned} \mathcal{L}_{\text{GCL}}^{(2)} = & - \sum_{u \in \mathcal{V}^{\text{DR}}} \log \frac{\exp((\bar{\mathbf{h}}_u^{\text{DR}})^T \mathbf{h}_{\text{HIN},u}^{\text{DR}})/\tau}{\sum_{j \in \{\mathcal{V}^{\text{DR}} - \mathcal{N}_{\mathcal{G}_{\text{DR}}}(u)\}} (\exp((\bar{\mathbf{h}}_u^{\text{DR}})^T \mathbf{h}_{\text{HIN},j}^{\text{DR}})/\tau)} \\ & - \sum_{v \in \mathcal{V}^{\text{DI}}} \log \frac{\exp((\bar{\mathbf{h}}_v^{\text{DI}})^T \mathbf{h}_{\text{HIN},v}^{\text{DI}})/\tau}{\sum_{j \in \{\mathcal{V}^{\text{DI}} - \mathcal{N}_{\mathcal{G}_{\text{DI}}}(v)\}} (\exp((\bar{\mathbf{h}}_v^{\text{DI}})^T \mathbf{h}_{\text{HIN},j}^{\text{DI}})/\tau)} \end{aligned} \quad (11)$$

### Adaptive gradient balance

This module aims to dynamically adapt the gradient of GCL auxiliary tasks from the perspective of gradient direction and magnitude for better transferring knowledge to main task.

Before delving into the details of our proposed strategy, we start with briefly introducing the general optimization procedure of existing GCL-based DDA studies. Conventionally, the GCL auxiliary tasks are jointly optimized along with the main task, i.e. DDA prediction. Formally, let  $\theta$  denote the shared parameters, the multi-task loss function is defined as  $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{DDA}} + \sum_{i=1}^N \mathcal{L}_{\text{GCL}}^{(i)}$ , where  $N$  means the total number of auxiliary tasks. The corresponding gradient of  $\mathcal{L}_{\text{total}}$  w.r.t.  $\theta$  at  $t$ -th training iteration is

calculated as follows:

$$\mathbf{G}_{\text{total}}^t = \mathbf{G}_{\text{DDA}}^t + \sum_{i=1}^N \mathbf{G}_{\text{GCL}}^{(i)t} = \nabla_{\theta} \mathcal{L}_{\text{DDA}}^t + \sum_{i=1}^N \nabla_{\theta} \mathcal{L}_{\text{GCL}}^{(i)t} \quad (12)$$

we assume that  $\theta$  is updated by gradient descent with the learning rate  $\alpha$ , i.e.,  $\theta^{t+1} = \theta^t - \alpha \times \mathbf{G}_{\text{total}}^t$ . Hence, if  $\|\mathbf{G}_{\text{GCL}}^{(i)t}\| \gg \|\mathbf{G}_{\text{DDA}}^t\|$  ( $\|\cdot\|$  is  $L_2$  norm), the optimizer inclines to update shared parameters  $\theta$  toward the  $i$ -th GCL auxiliary task rather than main task, resulting in serious issue of optimization imbalance. A straightforward solution is to add weight hyperparameters for each auxiliary task, i.e.  $\mathbf{G}_{\text{total}}^t = \mathbf{G}_{\text{DDA}}^t + \sum_{i=1}^N \lambda_i \mathbf{G}_{\text{GCL}}^{(i)t}$ . However, as mentioned by [16, 18], task hyperparameters retain a linear increase with the number of defined auxiliary tasks, and tuning the hyperparameters for multiple GCL tasks is burdensome. More importantly, since gradient magnitudes are dynamically changing during the entire optimization process, such fixed task hyperparameters trap in local optimum.

In this paper, our goal is to balance  $\mathbf{G}_{\text{DDA}}^t$  and  $\mathbf{G}_{\text{GCL}}^{(i)t}$  at each epoch through adaptively adjusting the direction and magnitude of  $\mathbf{G}_{\text{GCL}}^{(i)t}$ . Specifically, the moving average of gradient magnitude is utilized to take into account the variance over training iterations [30]:

$$\begin{aligned} m_{\text{DDA}}^t &= \beta \times m_{\text{DDA}}^{t-1} + (1 - \beta) \times \|\mathbf{G}_{\text{DDA}}^t\| \\ m_{\text{GCL}}^{(i)t} &= \beta \times m_{\text{GCL}}^{(i)t-1} + (1 - \beta) \times \|\mathbf{G}_{\text{GCL}}^{(i)t}\| \end{aligned} \quad (13)$$

where  $\beta$  is a hyperparameter,  $m_{\text{DDA}}^0$  and  $m_{\text{GCL}}^{(i)0}$  are both initialized with 0. When  $m_{\text{GCL}}^{(i)t} > m_{\text{DDA}}^t$ , inspired by prior works [17, 18], we firstly alter the gradient direction by projecting the gradient of  $i$ -th auxiliary task  $\mathbf{G}_{\text{GCL}}^{(i)t}$  to the normal plane of main task gradient  $\mathbf{G}_{\text{DDA}}^t$ , if these two gradients conflict with each other ( $\cosine(\mathbf{G}_{\text{DDA}}^t, \mathbf{G}_{\text{GCL}}^{(i)t}) < 0$ ). The gradient direction modification is formulated as follows:

$$\mathbf{G}_{\text{GCL}}^{(i)t} = \mathbf{G}_{\text{GCL}}^{(i)t} - \frac{\mathbf{G}_{\text{GCL}}^{(i)t} \cdot \mathbf{G}_{\text{DDA}}^t}{\|\mathbf{G}_{\text{DDA}}^t\|^2} \mathbf{G}_{\text{DDA}}^t \quad (14)$$

where  $\cdot$  refers to dot product. Please note that when  $m_{\text{GCL}}^{(i)t} \leq m_{\text{DDA}}^t$ , even through these two gradient directions are conflicting,  $\mathbf{G}_{\text{GCL}}^{(i)t}$  remains the same as before, thereby preventing overfitting.

Next, since larger gradient magnitude dominates the overall optimization trend, we further alter the magnitude proximity between  $\mathbf{G}_{\text{GCL}}^{(i)t}$  and  $\mathbf{G}_{\text{DDA}}^t$  by introducing a relax factor  $\omega$  to enhance the flexibility of magnitude scaling [16]:

$$\mathbf{G}_{\text{GCL}}^{(i)t} = \omega \times \frac{\|\mathbf{G}_{\text{DDA}}^t\|}{\|\mathbf{G}_{\text{GCL}}^{(i)t}\|} \mathbf{G}_{\text{GCL}}^{(i)t} + (1 - \omega) \times \mathbf{G}_{\text{GCL}}^{(i)t} \quad (15)$$

Through the above process, we succeed in adjusting the gradient of GCL auxiliary tasks at each training iteration from the perspective of gradient direction and magnitude. The overall procedure of parameter update is summarized in Algorithm 1. It is worth noting that the moving average coefficient  $\beta$  is empirically set to 0.9. Thus, just one hyperparameter, i.e. relax factor  $\omega$ , needs to be tuned, irrespective of the number of GCL auxiliary tasks. Furthermore, time complexity analysis of GCGB can be found in the supplementary material.

## Experiments

In this section, we first outline the experimental setup in Section Experimental Setup. After that, we compare GCGB with competitive baselines in Section Main Results and Analyses. The head-to-head comparison between GCGB and a baseline is discussed in Section Head-to-head Comparison. Subsequently, the ablation studies are provided in Section Ablation Studies. Section Robustness against Label Sparsity discusses the robustness analysis against label sparsity. The effectiveness of adaptive gradient balance is then described in Section Effectiveness of Adaptive Gradient Balance. Finally, detailed case studies to evaluate the performance consistency among different disease and drug categories, and further predict candidate drugs for two neurodegenerative diseases are presented in Section Case Studies. More carefully designed experiments, such as cold-start scenario and generalization evaluation etc., are available in the supplementary material.

### Experimental setup

#### Evaluation metrics

To validate the prediction performance of GCGB, seven evaluation metrics are adopted, including area under the receiver operating characteristic (ROC) curve (AUC), area under the precision-recall (PR) curve (AUPR), accuracy, precision, recall, F1-score, and Matthews correlation coefficient (MCC). For all evaluation metrics, higher scores indicate better performance.

#### Implementation details

Since all the benchmark datasets solely comprise positive DDAs, we firstly generate negative samples through randomly pairing the drugs and diseases which have unconfirmed associations. Hence, these datasets are carefully balanced to ensure equal number of positive and negative samples. Moreover, to avoid the bias of experimental results, 10-fold cross-validation is employed to evaluate the predictive performance.

When constructing the drug and disease similarity networks, the nearest neighbor number  $K$  is set to 20. Besides, throughout our experiments, we set the node embedding dimension  $d = 256$ . As for dual GT networks, the number of layers and heads are 2 and 4, respectively. The temperature coefficient  $\tau$  in Equations 10-11 is tuned among {0.05, 0.2, 0.5, 1} (The parameter sensitivity analysis is provided in supplementary material). The trade-off parameters  $\beta$  (in Equation 13) and  $\omega$  (in Equation 15) are both empirically set to 0.9. The overall model parameters are initialized with Xavier and are optimized by Adam with an initial learning rate  $\alpha = 0.0002$ .

#### Baselines

The competitive baselines are categorized into the following four groups: (i) deep learning (DL)-based method, including HNet-DNN [31]; (ii) graph representation learning (GRL)-based methods, including HINGRL [32], RLFDDA [33], and SFRLDDA [1]; (iii) GNN-based methods, including DRHGCN [34], DDAGDL [35], DRWBNCF [36], and AMDGT [7]; (iv) GCL-based methods, including SGCD [5], SADR [12], and DRGCL [37]. The elaborate descriptions of baselines are provided in supplementary material.

### Main results and analyses

As shown in Table 2, our proposed GCGB is substantially superior to all the competitive baselines on three benchmarks. GCGB also presents statistically significant improvements in terms of most evaluation metrics (with Welch's t-test P-value  $< 0.05$  or P-value  $< 0.01$ ). It is worth noting that the precision scores

Table 2. Test results compared with baselines on three benchmark datasets

Dataset	Model	Evaluation criteria						
		AUC	AUPR	Accuracy	Precision	Recall	F1	MCC
B-dataset	HNet-DNN	0.8927 $\pm$ 0.002	0.8919 $\pm$ 0.001	0.8101 $\pm$ 0.001	0.7825 $\pm$ 0.001	0.8281 $\pm$ 0.002	0.8047 $\pm$ 0.001	0.6211 $\pm$ 0.002
	HINGRL	0.8845 $\pm$ 0.003	0.8774 $\pm$ 0.002	0.8035 $\pm$ 0.002	0.8006 $\pm$ 0.003	0.8084 $\pm$ 0.004	0.8045 $\pm$ 0.004	0.6071 $\pm$ 0.004
	RLFDDA	0.8728 $\pm$ 0.006	–	0.7907 $\pm$ 0.006	0.7821 $\pm$ 0.008	0.8060 $\pm$ 0.008	0.7938 $\pm$ 0.006	–
	SFRLDDA	0.8364 $\pm$ 0.280	0.8271 $\pm$ 0.490	0.7588 $\pm$ 0.600	0.7503 $\pm$ 0.610	0.7757 $\pm$ 0.770	0.7628 $\pm$ 0.600	0.5178 $\pm$ 1.200
	DRHGCN	0.9092 $\pm$ 0.002	0.9106 $\pm$ 0.002	0.8268 $\pm$ 0.002	0.8678 $\pm$ 0.001	0.7711 $\pm$ 0.001	0.8166 $\pm$ 0.001	0.6577 $\pm$ 0.001
	DDAGDL	0.8421 $\pm$ 0.003	0.8315 $\pm$ 0.002	0.7646 $\pm$ 0.003	0.7616 $\pm$ 0.004	0.7703 $\pm$ 0.002	0.7659 $\pm$ 0.004	0.5292 $\pm$ 0.003
	DRWBNCf	0.9004 $\pm$ 0.001	0.9018 $\pm$ 0.002	0.5991 $\pm$ 0.002	<b>0.9810 <math>\pm</math> 0.002</b>	0.2021 $\pm$ 0.004	0.3352 $\pm$ 0.003	0.3260 $\pm$ 0.003
	AMDGT	<u>0.9317 <math>\pm</math> 0.002</u>	<u>0.9302 <math>\pm</math> 0.003</u>	<u>0.8593 <math>\pm</math> 0.002</u>	0.8612 $\pm$ 0.003	0.8619 $\pm$ 0.002	<u>0.8616 <math>\pm</math> 0.003</u>	<u>0.7215 <math>\pm</math> 0.003</u>
	SGCD	0.9235 $\pm$ 0.003	0.9224 $\pm$ 0.002	0.8506 $\pm$ 0.004	0.8458 $\pm$ 0.005	<u>0.8631 <math>\pm</math> 0.003</u>	0.8536 $\pm$ 0.003	0.7018 $\pm$ 0.004
	SADR	0.9211 $\pm$ 0.004	0.9209 $\pm$ 0.003	0.8525 $\pm$ 0.005	0.8512 $\pm$ 0.003	0.8599 $\pm$ 0.004	0.8548 $\pm$ 0.003	0.7046 $\pm$ 0.005
	DRGCL	0.9247 $\pm$ 0.003	0.9230 $\pm$ 0.005	0.8546 $\pm$ 0.006	0.8544 $\pm$ 0.008	0.8626 $\pm$ 0.007	0.8582 $\pm$ 0.006	0.7133 $\pm$ 0.008
	GCGB (Ours)	<b>0.9369 <math>\pm</math> 0.004**</b>	<b>0.9344 <math>\pm</math> 0.004*</b>	<b>0.8766 <math>\pm</math> 0.008**</b>	<u>0.8728 <math>\pm</math> 0.011</u>	<b>0.8826 <math>\pm</math> 0.009**</b>	<b>0.8773 <math>\pm</math> 0.008**</b>	<b>0.7534 <math>\pm</math> 0.015**</b>
C-dataset	HNet-DNN	0.9460 $\pm$ 0.002	0.9399 $\pm$ 0.001	0.8838 $\pm$ 0.001	0.8778 $\pm$ 0.002	0.8820 $\pm$ 0.001	0.8799 $\pm$ 0.001	0.7674 $\pm$ 0.002
	HINGRL	0.9372 $\pm$ 0.004	0.9457 $\pm$ 0.005	0.8698 $\pm$ 0.002	0.8851 $\pm$ 0.004	0.8500 $\pm$ 0.004	0.8672 $\pm$ 0.003	0.7403 $\pm$ 0.002
	RLFDDA	0.9636 $\pm$ 0.005	–	0.9006 $\pm$ 0.012	0.9035 $\pm$ 0.014	0.8972 $\pm$ 0.022	0.9002 $\pm$ 0.013	–
	SFRLDDA	0.9519 $\pm$ 0.520	0.9586 $\pm$ 0.320	0.8934 $\pm$ 1.290	0.8824 $\pm$ 1.630	0.9080 $\pm$ 1.750	0.8949 $\pm$ 1.260	0.7873 $\pm$ 2.580
	DRHGCN	0.9324 $\pm$ 0.003	0.9427 $\pm$ 0.004	0.8652 $\pm$ 0.002	<b>0.9192 <math>\pm</math> 0.001</b>	0.8008 $\pm$ 0.002	0.8559 $\pm$ 0.002	0.7366 $\pm$ 0.003
	DDAGDL	0.8693 $\pm$ 0.003	0.8935 $\pm$ 0.004	0.8168 $\pm$ 0.002	0.7874 $\pm$ 0.004	0.7721 $\pm$ 0.002	0.7797 $\pm$ 0.003	0.6230 $\pm$ 0.003
	DRWBNCf	0.9234 $\pm$ 0.004	0.9419 $\pm$ 0.004	0.8663 $\pm$ 0.004	0.8984 $\pm$ 0.002	0.8370 $\pm$ 0.004	0.8612 $\pm$ 0.004	0.7449 $\pm$ 0.003
	AMDGT	<u>0.9672 <math>\pm</math> 0.003</u>	<u>0.9696 <math>\pm</math> 0.003</u>	<u>0.9052 <math>\pm</math> 0.003</u>	0.8912 $\pm$ 0.003	<u>0.9250 <math>\pm</math> 0.004</u>	<u>0.9078 <math>\pm</math> 0.003</u>	<u>0.8122 <math>\pm</math> 0.004</u>
	SGCD	0.9564 $\pm$ 0.004	0.9555 $\pm$ 0.006	0.8974 $\pm$ 0.003	0.8658 $\pm$ 0.005	0.8717 $\pm$ 0.003	0.8695 $\pm$ 0.003	0.7432 $\pm$ 0.004
	SADR	0.9550 $\pm$ 0.005	0.9584 $\pm$ 0.003	0.8994 $\pm$ 0.005	0.8768 $\pm$ 0.003	0.8853 $\pm$ 0.004	0.8806 $\pm$ 0.004	0.7683 $\pm$ 0.003
	DRGCL	0.9606 $\pm$ 0.003	0.9619 $\pm$ 0.004	0.9041 $\pm$ 0.006	0.8988 $\pm$ 0.005	0.9045 $\pm$ 0.006	0.9013 $\pm$ 0.008	0.8077 $\pm$ 0.010
	GCGB (Ours)	<b>0.9713 <math>\pm</math> 0.004*</b>	<b>0.9746 <math>\pm</math> 0.004**</b>	<b>0.9226 <math>\pm</math> 0.009**</b>	<u>0.9162 <math>\pm</math> 0.013</u>	<b>0.9305 <math>\pm</math> 0.012**</b>	<b>0.9232 <math>\pm</math> 0.008**</b>	<b>0.8455 <math>\pm</math> 0.018**</b>
F-dataset	HNet-DNN	0.9188 $\pm$ 0.002	0.9157 $\pm$ 0.001	0.8426 $\pm$ 0.002	0.8502 $\pm$ 0.002	0.8413 $\pm$ 0.002	0.8457 $\pm$ 0.001	0.6851 $\pm$ 0.001
	HINGRL	0.9366 $\pm$ 0.006	0.9449 $\pm$ 0.004	0.8645 $\pm$ 0.005	0.8832 $\pm$ 0.004	0.8402 $\pm$ 0.003	0.8612 $\pm$ 0.006	0.7300 $\pm$ 0.004
	SFRLDDA	0.9164 $\pm$ 0.640	0.9266 $\pm$ 0.810	0.8414 $\pm$ 1.300	0.8345 $\pm$ 1.340	0.8520 $\pm$ 2.430	0.8430 $\pm$ 1.390	0.6834 $\pm$ 2.620
	DRHGCN	0.9207 $\pm$ 0.004	0.9375 $\pm$ 0.002	0.8583 $\pm$ 0.001	<b>0.9309 <math>\pm</math> 0.001</b>	0.7739 $\pm$ 0.002	0.8452 $\pm$ 0.002	0.7269 $\pm$ 0.002
	DDAGDL	0.9239 $\pm$ 0.007	0.9235 $\pm$ 0.002	0.8513 $\pm$ 0.004	0.8475 $\pm$ 0.005	0.8567 $\pm$ 0.004	0.8521 $\pm$ 0.005	0.7026 $\pm$ 0.003
	DRWBNCf	0.8958 $\pm$ 0.005	0.9200 $\pm$ 0.004	0.8296 $\pm$ 0.002	0.8752 $\pm$ 0.003	0.8237 $\pm$ 0.004	0.8341 $\pm$ 0.004	0.7232 $\pm$ 0.002
	AMDGT	<u>0.9584 <math>\pm</math> 0.005</u>	<u>0.9607 <math>\pm</math> 0.003</u>	0.8908 $\pm$ 0.003	0.8730 $\pm$ 0.003	0.9146 $\pm$ 0.003	0.8928 $\pm$ 0.005	0.7815 $\pm$ 0.004
	SGCD	0.9496 $\pm$ 0.003	0.9550 $\pm$ 0.004	0.8940 $\pm$ 0.004	0.8663 $\pm$ 0.005	<u>0.9175 <math>\pm</math> 0.005</u>	0.8894 $\pm$ 0.004	0.7711 $\pm$ 0.004
	SADR	0.9504 $\pm$ 0.006	0.9578 $\pm$ 0.005	0.8914 $\pm$ 0.003	0.8719 $\pm$ 0.005	0.9123 $\pm$ 0.004	0.8916 $\pm$ 0.005	0.7785 $\pm$ 0.004
	DRGCL	0.9525 $\pm$ 0.005	0.9602 $\pm$ 0.004	<u>0.8985 <math>\pm</math> 0.007</u>	0.8879 $\pm$ 0.008	0.9152 $\pm$ 0.006	<u>0.9005 <math>\pm</math> 0.008</u>	<u>0.7843 <math>\pm</math> 0.009</u>
	GCGB (Ours)	<b>0.9676 <math>\pm</math> 0.006**</b>	<b>0.9714 <math>\pm</math> 0.006**</b>	<b>0.9219 <math>\pm</math> 0.009**</b>	<u>0.9152 <math>\pm</math> 0.011</u>	<b>0.9364 <math>\pm</math> 0.013**</b>	<b>0.9230 <math>\pm</math> 0.010**</b>	<b>0.8445 <math>\pm</math> 0.021**</b>

The reported results are in the form of (mean  $\pm$  standard deviation) under 10-fold cross-validation. For fair comparison, all the baselines are compared under identical evaluation settings. Specifically, as for the approaches with publicly released codes, we reproduce the experiments according to the best parameters recommended in their original papers. For those methods without released codes, the predictive performances are directly taken from the original papers, “–” indicates that the corresponding experimental result is not reported in previous works. The best score is in **bold**, and the second best score is underlined. “\*” and “\*\*” denote that our proposed method significantly outperforms the strongest baseline at corresponding metric based on Welch’s t-test (“\*” refers to  $P$ -value  $< 0.05$ , “\*\*” refers to  $P$ -value  $< 0.01$ ).

obtained by DRWBNCf and DRHGCN are much higher than their recall scores, indicating that these methods are prone to identify the known DDAs as negatives, whereas our approach exhibits slighter performance fluctuation across all metrics.

In the following, we will compare GCGB with each group of baselines in turn and discuss the superiority of our approach. Firstly, the performance of DL-based method, i.e. HNet-DNN is relatively mediocre, since it extracts drug and disease features from Euclidean space, while neglects the non-Euclidean geometric property [35].

Secondly, compared with GRL-based methods, which apply different representation learning strategies to obtain the embeddings of drugs and diseases from both similarity and association networks, GCGB achieves far better performance. We argue that GRL-based baselines neglect the crucial neighborhood information which is conducive to producing discriminative representations for drugs and diseases, leading to the compromised effectiveness.

Thirdly, to adequately aggregate neighbors’ messages within graph structures, GNN-based methods are proposed. We observe that GCGB significantly outperforms these methods, the main reason lies in their inability to relieve the issue of label sparsity, resulting in ineffectively capturing intricate semantic correlations between drugs and diseases. On the contrary, in this paper, self-supervised GCL auxiliary tasks are explicitly introduced to combat against the label scarcity and further enhance the representation capacity.

Finally, GCGB consistently prevails over GCL-based baselines, indicating the benefits of inter-view CL and adaptive gradient balance. Specifically, SGCD and SADR either utilize representational or structural perturbation strategies to construct augmented views. However, blindly corrupting graph topological structures causes the absence of necessary associations between drugs and diseases, thereby leading to performance degradation [13]. Moreover, even if DRGCL defines an inter-view CL auxiliary task by aligning topology and semantic information, thus obviating

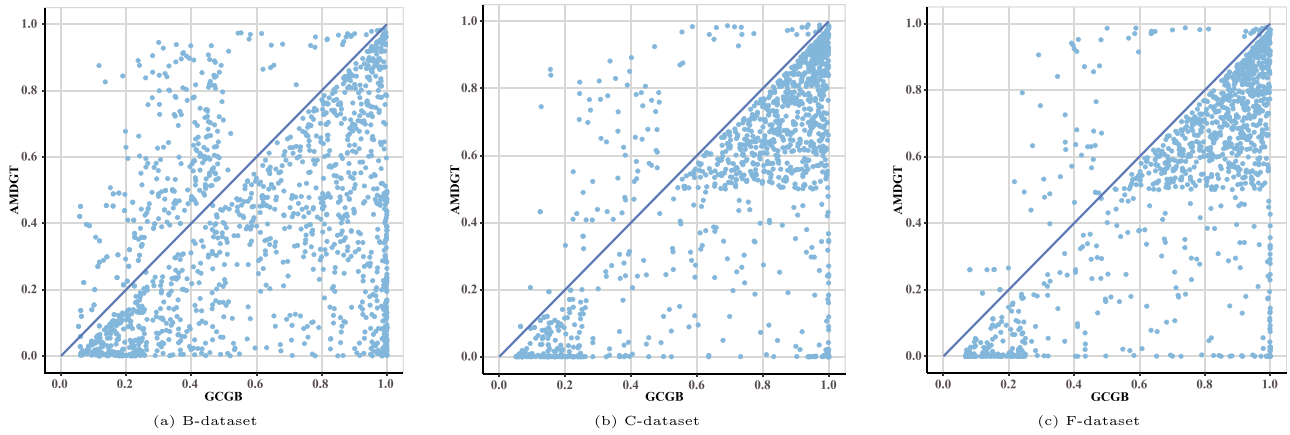


Figure 2. Head-to-head comparison between GCGB and the strongest baseline AMDGT on three benchmarks.

the need for input data augmentation, it still ignores the imbalance phenomenon of gradient direction and magnitude between GCL auxiliary tasks and main task. To bridge the gap, in this paper, we innovatively contrast the fusion view with semantic and interaction views, respectively, to capture higher-order interaction semantics. Moreover, the optimization imbalance phenomenon between GCL auxiliary tasks and main task is significantly alleviated to improve predictive performance.

### Head-to-head comparison

The DDA prediction is formulated as a binary classification task. Intuitively, the higher the predicted scores for positive class, the better the model performs. Therefore, we take a step further to conduct the head-to-head comparison between our proposed GCGB and the strongest baseline AMDGT (please refer to Table 2). For better visualization, following the setting in [38], we subsample the labeled positive DDAs down to 1000 samples, the performance comparison results on three datasets are shown in scatter plots of Fig. 2, in which each point denotes a DDA sample, and the x-axis refers to the output results of our model, while the y-axis represents the predicted scores of AMDGT. We observe that the majority of data points are located below the diagonal, suggesting that GCGB not only correctly predicts the label of positive DDAs, but also consistently assigns higher confidence scores than AMDGT. It is worth noting that those points scattered above the diagonal do not necessarily indicate wrong predictions, and the points gathered in the lower left corner are hard samples that are difficult to predict for both methods.

### Ablation studies

To evaluate the effect of different components within GCGB, we conduct model ablation studies over several variants under 10-fold cross-validation. The average results are reported. As presented in Table 3, the ablation results manifest that each component contributes to the final performance.

Firstly, we replace the HGT with basic GT when modeling the interaction view in Section Interaction View Feature Extraction. From the results, ablated model performs worse than GCGB, indicating that HGT is capable of sufficiently encoding informative interaction patterns and capturing heterogeneous properties within interaction view.

Secondly, we remove the adaptive gradient balance module introduced in Section Adaptive Gradient Balance. In other words,

we do not adjust the gradient of GCL auxiliary tasks at each training iteration. Unfortunately, the corresponding ablation performance decreases by a certain margin, clearly revealing that the optimization imbalance phenomenon between auxiliary and main tasks is detrimental to the predictive performance (in-depth analysis about the effectiveness of adaptive gradient balance is provided in Section Effectiveness of Adaptive Gradient Balance).

Thirdly, we simultaneously ignore the inter-view contrastive loss between fusion view and semantic views (Equation 10) as well as the adaptive gradient balance module. The corresponding ablation result drops dramatically, which shows the significance of this inter-view CL auxiliary task.

Finally, we also jointly remove the inter-view contrastive loss between fusion view and interaction view (Equation 11) as well as the adaptive gradient balance module. As expected, this ablation variant results in a drastic performance drop. Above observations demonstrate that the fusion view could integrate both higher-order relationships (drug-drug and disease-disease) and interaction information (drug-disease). Moreover, these two inter-view CL auxiliary tasks are effective in learning high-quality drug and disease representations through supplementing additional training pseudo-labels.

### Robustness against label sparsity

To mitigate the effect of label sparsity encountered by GNN-based methods, we contrast the fusion view with semantic and interaction views, respectively. Herein, we aim to illustrate the effectiveness of GCGB in alleviating label sparsity. Specifically, we separate the diseases into five groups according to their sparsity degrees, i.e. the number of known associations with drugs, and report the comparison results with AMDGT under 10-fold cross-validation in Fig. 3. Based on the results, we have the following observations: (i) The sparsity degree of diseases exhibits a clear long-tail distribution, i.e. a significant majority of diseases have few interactions with drugs (please refer to the dark blue region at the bottom of stacked bar chart). (ii) Compared with AMDGT, GCGB consistently achieves better AUC and AUPR performance on three datasets, especially for highly sparse diseases, indicating the robustness of GCGB in handling label scarcity. Moreover, We further study the robustness of GCGB against label sparsity by evaluating the model performance on different sparsity level of training data. Please refer to Section S3.2 *Performance on Sparsified Datasets* in the supplementary material for details.



Table 3. Ablation studies on three benchmark datasets

Dataset	Model	Evaluation criteria			
		AUC	AUPR	F1	MCC
B-dataset	GCGB (entire model)	<b>0.9369</b>	<b>0.9344</b>	<b>0.8773</b>	<b>0.7534</b>
	w/o HGT	0.9340	0.9297	0.8697	0.7340
	w/o Gradient balance	0.9278	0.9247	0.8654	0.7281
	w/o Semantic view contrastive	0.9073	0.9057	0.8427	0.6942
	w/o Interaction view contrastive	0.9188	0.9148	0.8487	0.7034
C-dataset	GCGB (entire model)	<b>0.9713</b>	<b>0.9746</b>	<b>0.9232</b>	<b>0.8455</b>
	w/o HGT	0.9653	0.9658	0.9139	0.8270
	w/o Gradient balance	0.9614	0.9642	0.9097	0.8226
	w/o Semantic view contrastive	0.9413	0.9461	0.8895	0.7885
	w/o Interaction view contrastive	0.9578	0.9584	0.8990	0.8014
F-dataset	GCGB (entire model)	<b>0.9676</b>	<b>0.9714</b>	<b>0.9230</b>	<b>0.8445</b>
	w/o HGT	0.9626	0.9664	0.9104	0.8208
	w/o Gradient balance	0.9596	0.9639	0.9036	0.8084
	w/o Semantic view contrastive	0.9257	0.9342	0.8879	0.7740
	w/o Interaction view contrastive	0.9334	0.9381	0.8947	0.7927

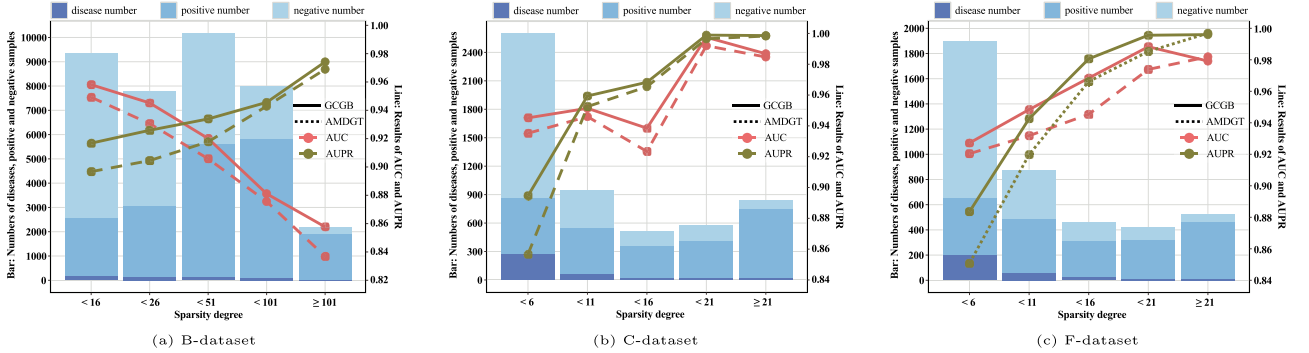


Figure 3. The stacked bar chart displays the corresponding number of diseases, positive and negative training samples partitioned by disease sparsity degree. The line chart shows the corresponding AUC and AUPR results of GCGB and AMDGT w.r.t. different disease sparsity degree.

## Effectiveness of adaptive gradient balance

In the ablation studies, we have demonstrated that removing the adaptive gradient balance module indeed leads to the performance degradation (w/o Gradient balance in Table 3). Herein, taking C-dataset as an example, we will present in-depth analysis about the effectiveness of adaptive gradient balance from the perspective of gradient direction and magnitude.

As for gradient direction, Fig. 4a and b depicts the proportion of conflicting gradient directions between two inter-view CL auxiliary tasks and main task at each training epoch without (or with) our proposed adaptive gradient balance module. It is obvious that the conflicting proportion is drastically reduced by performing optimization balance. Please note that as introduced in Section Adaptive Gradient Balance, in order to avoid overfitting, we rectify the gradient direction of  $i$ -th auxiliary task only if  $m_{GCL}^{(i)t} > m_{DDA}^t$ .

Besides, as for gradient magnitude, Fig. 4c and d displays the change of gradient magnitudes during the entire training epochs without (or with) the adaptive gradient balance. The visualization results in Fig. 4c lead us to the following observations. Firstly, gradient magnitudes of GCL auxiliary tasks are much larger than the counterpart of main task. Secondly, the gradient magnitudes are dynamically changing during the training process. Above

observations highlight the gradient magnitude imbalance between GCL auxiliary tasks and main task. To alleviate this intractable issue, we innovatively perform magnitude scaling for each auxiliary task. Hence, as shown in Fig. 4d, the magnitudes of  $G_{GCL}^1$  and  $G_{GCL}^2$  are in close proximity to the counterpart of  $G_{DDA}$ , thereby preventing GCL auxiliary tasks from dominating the optimization process.

## Case studies

### Performance consistency among different disease and drug categories

Herein, we conduct systematic and unbiased validations to evaluate the predictive performance of GCGB among different disease and drug categories. Specifically, the corresponding experiments are carried out on the B-dataset. The reasons why we select B-dataset are listed below: (i) The DDAs number on B-dataset is an order of magnitude more than the ones of other two datasets (please refer to Table 1 for details). (ii) Since GCGB performs relatively worse on B-dataset than on other two datasets, it is more meaningful to verify the performance consistency on B-dataset (please refer to Table 2 for details). (iii) Last but not least, the diseases in B-dataset are collected from Comparative Toxicogenomics Database (CTD) database

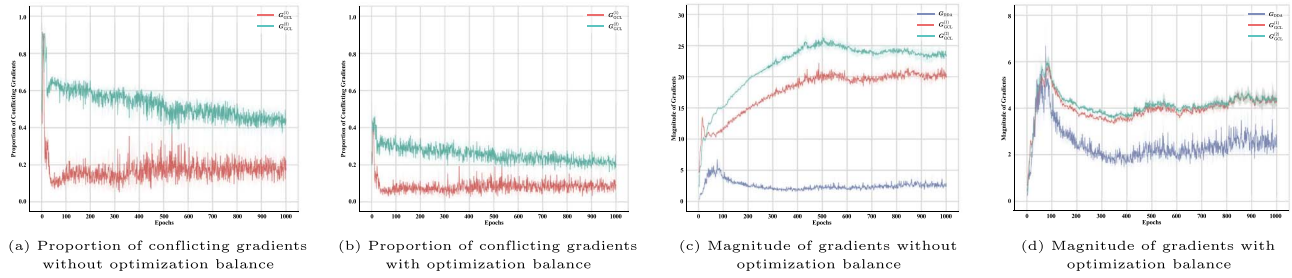


Figure 4. The proportion of conflicting gradient directions and the magnitude of gradients between the GCL auxiliary tasks and main task at each training epoch on C-dataset under 10-fold cross-validation. The solid lines stand for the average result, the upper and lower range represent 95% confidence interval.

Table 4. Predictive performance w.r.t. different disease categories

Category	Evaluation criteria			
	AUC	AUPR	F1	MCC
Animal diseases	0.9710	0.9622	0.8708	0.7906
Cardiovascular diseases	0.9233	0.9285	0.8797	0.7361
Chemically-induced disorders	0.9445	0.9541	0.9098	0.7911
Congenital, hereditary, neonatal diseases	0.9474	0.9207	0.8672	0.7832
Digestive system diseases	0.9427	0.9449	0.8858	0.7579
Endocrine system diseases	0.9630	0.9437	0.8852	0.8232
Eye diseases	0.9470	0.9183	0.8497	0.7715
Hemic and lymphatic diseases	0.9485	0.9351	0.8797	0.7840
Immune system diseases	0.9376	0.9236	0.8718	0.7750
Infections	0.9701	0.9436	0.7913	0.7380
Mental disorders	0.9354	0.9401	0.8865	0.7566
Musculoskeletal diseases	0.9324	0.9097	0.8370	0.7293
Neoplasms	0.9628	0.9511	0.8842	0.8121
Nervous system diseases	0.9331	0.9337	0.8764	0.7483
Nutritional and metabolic diseases	0.9395	0.9223	0.8568	0.7559
Otorhinolaryngologic diseases	0.9213	0.9042	0.8131	0.7150
Pathological conditions, signs and symptoms	0.9294	0.9390	0.8820	0.7337
Respiratory tract diseases	0.9268	0.9213	0.8481	0.7359
Skin and connective tissue diseases	0.9318	0.9260	0.8721	0.7548
Stomatognathic diseases	0.9441	0.9585	0.9043	0.7802
Urogenital diseases	0.9380	0.9314	0.8722	0.7627
Wounds and injuries	0.9648	0.9386	0.8663	0.7992

(<https://ctdbase.org/>), and the diseases in other two datasets are derived from Online Mendelian Inheritance in Man (OMIM) database. As for C-dataset and F-dataset, in order to obtain the information about disease categories, we map the OMIM ID of each disease to the International Classification of Diseases 10th Revision (ICD-10) code through OMIM website(<https://www.omim.org/>), whereas only 29.1% (119/409) and 29.7% (93/313) of the diseases in C-dataset and F-dataset can be converted to ICD-10 codes, respectively. Hence, we exclude these two datasets from our experiments.

As for B-dataset, the corresponding disease categories are generated by mapping disease names to the MEDIC-Slim classes, which are from the MeSH tree structures of disease branches. Moreover, to search the corresponding drug categories, we map the DrugBank ID of each drug to the 1st level of Anatomical Therapeutic Chemical (ATC) codes. Please note that one disease or drug can be assigned to multiple categories. The final numbers of disease and drug categories are 22 and 14, respectively. Table 4 and Table 5 present the predictive results of GCGB among different disease and drug categories. The consistent model performance is observed from above tables, revealing that GCGB has no predictive preference toward certain disease or drug categories. The

statistical data for each category are provided in the supplementary material.

### Drug prediction for neurodegenerative diseases

To further verify the predictive reliability of GCGB, following prior studies [7, 34, 39], we conduct detailed case studies to predict potential drugs for two neurodegenerative diseases, i.e. Alzheimer's and Parkinson's diseases (AD and PD for short) from the unknown DDAs within F-dataset.

AD is the most common dementing illness, which currently affects more than 55 million people worldwide. The specific brain abnormalities (amyloid- $\beta$  plaques and tau protein neurofibrillary tangles) influence the neurodegenerative process. However, the mechanisms leading to the accumulation of plaques as well as tangles are unknown, and removing amyloid- $\beta$  has not halted neurodegeneration [40]. There are no efficacious medications that have been licensed for use in individuals with AD [41]. Moreover, PD is a progressive neurodegenerative disorder typically characterized by the loss of dopaminergic neurons in the substantia nigra, which affects approximately 1–2% of the population aged 60 and older [42]. This disease presents with muscle stiffness, tremor, bradykinesia, and postural instability. Currently, the treatments of

Table 5. Predictive performance w.r.t. different drug categories

Category	Evaluation criteria			
	AUC	AUPR	F1	MCC
Alimentary tract and metabolism	0.9380	0.9420	0.8815	0.7551
Blood and blood forming organs	0.9407	0.9619	0.9139	0.7510
Cardiovascular system	0.9377	0.9381	0.8779	0.7503
Dermatologicals	0.9289	0.9381	0.8796	0.7290
Genito-urinary system and sex hormones	0.9221	0.9029	0.8363	0.7172
Systemic hormonal preparations	0.9328	0.9465	0.8880	0.7333
Antiinfectives for systemic use	0.9130	0.8988	0.8304	0.6929
Antineoplastic and immunomodulating agents	0.9351	0.9431	0.8983	0.7590
Musculo-skeletal system	0.9285	0.9303	0.8726	0.7346
Nervous system	0.9426	0.9496	0.9038	0.7748
Antiparasitic products, insecticides and repellents	0.9096	0.8800	0.8128	0.7043
Respiratory system	0.9383	0.9466	0.8889	0.7579
Sensory organs	0.9328	0.9465	0.8880	0.7333
Various	0.9504	0.9519	0.8999	0.7972

Table 6. Top-10 predicted drugs for Alzheimer's and Parkinson's diseases

Disease	Rank	Predicted drug	DrugBank ID	Researched or not
Alzheimer	1	Phenobarbital	DB01174	Yes
	2	Primidone	DB00794	No
	3	Cyproheptadine	DB00434	No
	4	Buspirone	DB00490	Yes
	5	Scopolamine	DB00747	Yes
	6	Citalopram	DB00215	Yes
	7	Doxorubicin	DB00997	No
	8	Fluoxetine	DB00472	Yes
	9	Haloperidol	DB00502	Yes
	10	Imipramine	DB00458	Yes
Parkinson	1	Buspirone	DB00490	Yes
	2	Biperiden	DB00810	Yes
	3	Risperidone	DB00734	Yes
	4	Carbamazepine	DB00564	No
	5	Rivastigmine	DB00989	Yes
	6	Clonazepam	DB01068	Yes
	7	Amantadine	DB00915	Yes
	8	Gabapentin	DB00996	Yes
	9	Primidone	DB00794	No
	10	Levodopa	DB01235	Yes

PD primarily revolve around alleviating symptoms and improving quality of life, there still lack established disease-modifying drugs [43]. In conclusion, due to a massive number of patients and the absence of effective therapeutic options, it is of significant clinical implications to identify potential candidate drugs for the above two neurodegenerative diseases.

Table 6 lists the top-10 candidate drugs discovered by GCGB for each disease and whether these predicted drugs have been studied for delaying disease progression by existing literatures (the evidences are available in supplementary material). Please note that AD and PD have no cure till now, the candidate drugs discovered by wet experimental-based and computational-based DR methods at most could delay the progression and alleviate symptoms. From the results in Table 6, we observe that as for AD, 7 out of 10 drugs have been researched in relevant medical literatures, and 8 predicted drugs have therapeutic potential for PD, indicating that GCGB is a trustworthy computational-based DR approach. However, the follow-up systematic preclinical

experiments and randomized clinical trials are required to verify the practical effects of predicted drugs. Besides, to further demonstrate the model generalization, following [35, 37, 44], we also predict top-10 potential drugs for breast cancer, please refer to the supplementary material for the results.

## Discussions

### Influence of protein targets

Considering that proteins play a variety of essential roles in living organisms, to enrich the graph connectivity, we integrate two protein-related bipartite graphs into the original DDA network, thereby composing the heterogeneous interaction network. However, plenty of disorders, e.g. infectious diseases caused by bacteria or viruses, are not associated with the up-regulation or down-regulation protein targets. Hence, we aim to examine whether GCGB can still achieve satisfactory performance when predicting

Table 7. Performance of GCGB when predicting DDAs for specific diseases without directly connected proteins

Dataset	Evaluation criteria			
	AUC	AUPR	F1	MCC
B-dataset	0.9393	0.9308	0.8734	0.7612
C-dataset	0.9724	0.9744	0.9240	0.8484
F-dataset	0.9648	0.9783	0.9266	0.8363

Table 8. Performance comparison on the virus-drug dataset

Model	Evaluation criteria	
	AUC	AUPR
DRRS	0.8214	0.8172
IRNMF	0.8122	0.7610
VAD	0.8372	0.8318
AntiViralDL	<u>0.8450</u>	<u>0.8494</u>
GCGB	<b>0.8821</b>	<b>0.8847</b>

DDAs for specific diseases with no directly connected protein targets. Specifically, we firstly collect all the diseases that have no first-order neighboring nodes of protein type on the interaction view  $\mathcal{G}_{\text{HIN}}$ . The numbers of such specific diseases are 456, 272, and 41 on the B-dataset, C-dataset, and F-dataset, respectively. Afterwards, all DDA pairs associated with above specific diseases are selected to evaluate the model performance. From the average results presented in Table 7, we observe that GCGB achieves consistent performance compared with the corresponding results listed in Table 2, and even gains slight performance improvement on several metrics, indicating that GCGB is not adversely affected by the absence of disease–protein interaction relationships. The major reason attributes to that the drug and disease representations are initially updated through two different views, i.e. semantic views and interaction view. Despite the absence of disease–protein interactions, the disease nodes could still aggregate information from their neighboring diseases and drugs from above two views, thereby refining the disease representations.

Furthermore, to analyze the generalization of GCGB for identifying antiviral drugs, we evaluate the model performance on a virus–drug association dataset constructed by [15], and compare GCGB with the competitive baselines, including DRRS [45], IRNMF [46], VAD [47], and AntiViralDL [15]. According to the reported results in [15], the average AUC and AUPR criteria are shown in

Table 8. We observe that GCGB surpasses all the baselines by a large margin, suggesting that GCGB is capable of discovering new antiviral drugs for clinical and biological research.

## Performance on extremely sparse scenarios

In the supplementary material (Section S3.2 *Performance on Sparsified Datasets*), we have evaluated the model performance on varying sparsity degrees of training data. Herein, we are interested in exploring the performance when applying GCGB to extremely sparse scenarios. Specifically, we merely sample 2, 1, and 0.5% training instances, and update the learnable parameters of GCGB on the extremely sparse labeled data. Through 10-fold cross-validation, the average results are recorded in Table 9. We observe that compared with the results obtained from entire datasets (Table 2), the corresponding performance on Table 9 drops by a large margin. The primary reason behind such failures is that even through the semantic view feature extraction submodule is not influenced by reducing the number of DDAs (semantic views are constructed according to multiple similarity measurement modalities), the extremely limited resources are insufficient for generating general and informative node representations from the interaction view, which further motivates our focus on zero-shot or few-shot DR, since 92% of total 17 080 diseases have no available medications, and up to 85% of rare diseases do not have even one developed drug [48]. We leave this exploration for future study.

## Conclusion

This paper proposes a novel heterogeneous GCL approach with gradient balance for inferring potential DDAs, namely GCGB. The primary innovations lie in the design of inter-view CL and adaptive gradient balance modules. Specifically, the fusion view is contrasted with semantic and interaction views respectively, thereby maximizing the mutual information between paired nodes across views. Furthermore, the gradients of GCL auxiliary tasks are dynamically adjusted from the perspective of gradient direction and magnitude for better guiding parameter update toward main task. Extensive experiments demonstrate that our approach consistently outperforms the competitive baselines on three commonly-used benchmarks under 10-fold cross-validation.

Regarding the future work, we would like to extend our research from the following two aspects. Firstly, due to the underlying black-box nature of artificial neural networks,

Table 9. Performance of GCGB on extremely sparse scenarios

Dataset	Data scale	Evaluation criteria			
		AUC	AUPR	F1	MCC
B-dataset	2%	0.7557	0.7502	0.7017	0.4494
	1%	0.7539	0.7473	0.6988	0.4473
	0.5%	0.7499	0.7423	0.6750	0.4327
C-dataset	2%	0.7747	0.7784	0.7396	0.4981
	1%	0.7720	0.7759	0.7132	0.4857
	0.5%	0.7616	0.7655	0.6276	0.4189
F-dataset	2%	0.7656	0.7539	0.6754	0.4701
	1%	0.7418	0.7322	0.5160	0.3747
	0.5%	0.7348	0.7303	0.4573	0.3246



the major limitation of GCGB lies in the lack of necessary transparency and interpretability, since the intermediate decision process is absent. We plan to investigate how to enhance the model interpretability through designing reinforcement learning-based models and formulating the DDA prediction as a sequential decision task. Secondly, in order to predict candidate drugs for diseases with limited or even no treatment options, we expect to transfer the implicit medical knowledge from well-annotated disorders to low-resource ones, such as rare diseases, via introducing an additional metric learning component.

#### Key Points

- A novel heterogeneous GCL method with gradient balance, namely GCGB, is proposed for inferring potential DDAs. To the best of our knowledge, it is the first time that optimization imbalance phenomenon between GCL auxiliary tasks and main task is considered in DR.
- We design effective inter-view CL auxiliary tasks through contrasting the fusion view with semantic and interaction views respectively, thereby maximizing the mutual information between paired nodes across views.
- To prevent auxiliary tasks from dominating the optimization process, we adaptively alter the gradient of GCL auxiliary tasks from the perspective of gradient direction and magnitude for better guiding parameter update toward main task.

## Acknowledgements

We would like to thank anonymous reviewers for their time and effort in reviewing this paper.

## Supplementary data

Supplementary data are available at Briefings in Bioinformatics online.

Conflict of interest: None declared.

## Funding

This work is supported by the National Natural Science Foundation of China under Grant 62072070 and the Fundamental Research Funds for the Central Universities under Grant 3132024265.

## Data availability

The benchmark datasets and the implementation of GCGB are available at [https://github.com/OleCui/paper\\_GCGB](https://github.com/OleCui/paper_GCGB).

## Author contributions

Hai Cui designed the research study. Hai Cui and Meiyu Duan wrote the manuscript and prepared figures. Haijia Bi conceived and conducted the experiments. Xiaobo Li and Xiaodi Hou analysed the results. Yijia Zhang supervised the research project. All authors contributed to proofreading and correcting the manuscript.

## References

1. Zhao B-W, Xiao-Rui S, Yang Y. et al. Drug-disease association prediction using semantic graph and function similarity representation learning over heterogeneous information networks. *Methods* 2023;**220**:106–14. <https://doi.org/10.1016/j.ymeth.2023.10.014>.
2. Tang J, Chen W, Zeng X. et al. GTDDA: graph convolutional network and graph transformer structure for drug repositioning. In: *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 436–9. Istanbul, Türkiye: IEEE, 2023.
3. He S, Yun L, Yi H. Fusing graph transformer with multi-aggregate GCN for enhanced drug-disease associations prediction. *BMC Bioinformatics* 2024;**25**:79. <https://doi.org/10.1186/s12859-024-05705-w>.
4. Mei X, Cai X, Yang L. et al. Relation-aware heterogeneous graph transformer based drug repurposing. *Expert Syst Appl* 2022;**190**:116165. <https://doi.org/10.1016/j.eswa.2021.116165>.
5. Huimin Y, Mingyu L, Li Z. et al. Semantic-enhanced graph contrastive learning with adaptive denoising for drug repositioning. *IEEE J Biomed Health Inform* 2023;**1**–9. <https://doi.org/10.1109/JBHI.2023.3344031>.
6. Zeng P, Zhang B, Liu A. et al. Drug repositioning based on tripartite cross-network embedding and graph convolutional network. *Expert Syst Appl* 2024;**252**:124152. <https://doi.org/10.1016/j.eswa.2024.124152>.
7. Liu J, Guan S, Zou Q. et al. AMDGT: attention aware multi-modal fusion using a dual graph transformer for drug–disease associations prediction. *Knowl -Based Syst* 2024;**284**:111329. <https://doi.org/10.1016/j.knosys.2023.111329>.
8. Sun X, Jia X, Zhangli L. et al. Drug repositioning with adaptive graph convolutional networks. *Bioinformatics* 2024;**40**:btad748. <https://doi.org/10.1093/bioinformatics/btad748>.
9. Liu B-M, Gao Y-L, Li F. et al. SLGCN: structure-enhanced line graph convolutional network for predicting drug–disease associations. *Knowl-Based Syst* 2024;**283**:111187. <https://doi.org/10.1016/j.knosys.2023.111187>.
10. Liu L, Huang F, Liu X. et al. Multi-view contrastive learning hypergraph neural network for drug-microbe-disease association prediction. In: *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 4829–37. Macao, SAR, China: IJCAI, 2023.
11. Wang X, Cheng Y, Yang Y. et al. Multitask joint strategies of self-supervised representation learning on biomedical networks for drug discovery. *Nat Mach Intell* 2023;**5**:445–56. <https://doi.org/10.1038/s42256-023-00640-6>.
12. Jin S, Zhang Y, Huimin Y. et al. SADR: self-supervised graph learning with adaptive denoising for drug repositioning. *IEEE/ACM Trans Comput Biol Bioinform* 2024;**21**:265–77. <https://doi.org/10.1109/TCBB.2024.3351079>.
13. Gao Z, Ma H, Zhang X. et al. Similarity measures-based graph co-contrastive learning for drug–disease association prediction. *Bioinformatics* 2023;**39**:btad357. <https://doi.org/10.1093/bioinformatics/btad357>.
14. Fan Y, Zhang C, Xiaowen H. et al. SGCLDGA: unveiling drug–gene associations through simple graph contrastive learning. *Brief Bioinform* 2024;**25**:bbae231. <https://doi.org/10.1093/bib/bbae231>.
15. Zhang P, Xiaowen H, Li G. et al. AntiViralDL: computational antiviral drug repurposing using graph neural network and self-supervised learning. *IEEE J Biomed Health Inform* 2023;**28**:548–56. <https://doi.org/10.1109/JBHI.2023.3328337>.
16. He Y, Xue F, Cheng C. et al. Metabalance: improving multi-task recommendations via adapting gradient magnitudes of

- auxiliary tasks. In: *Proceedings of the ACM Web Conference (WWW)*, pp. 2205–15. Lyon, France: ACM, 2022.
17. Tianhe Y, Kumar S, Gupta A. et al. Gradient surgery for multi-task learning. *Adv Neural Inf Process Syst (NeurIPS)* 2020;**33**: 5824–36.
  18. Xu J, Wang C, Wu C. et al. Multi-behavior self-supervised learning for recommendation. In: *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pp. 496–505. Taipei, Taiwan: ACM, 2023.
  19. Dwivedi VP, Bresson X. A generalization of transformer networks to graphs. *CoRR* abs/2012.09699 2020. <https://arxiv.org/abs/2012.09699>.
  20. Hu Z, Dong Y, Wang K. et al. Heterogeneous graph transformer. In: *Proceedings of the Web Conference (WWW)*, pp. 2704–10. Taipei, Taiwan, ACM, 2020.
  21. Vaswani A, Shazeer N, Parmar N. et al. Attention is all you need. *Adv Neural Inf Process Syst (NeurIPS)*, pp. 5998–6008. Long Beach, CA: USA, Curran Associates, Inc, 2017.
  22. Zhang W, Yue X, Lin W. et al. Predicting drug-disease associations by using similarity constrained matrix factorization. *BMC Bioinformatics* 2018;**19**:1–12. <https://doi.org/10.1186/s12859-018-2220-4>.
  23. Luo H, Wang J, Li M. et al. Drug repositioning based on comprehensive similarity measures and Bi-Random walk algorithm. *Bioinformatics* 2016;**32**:2664–71. <https://doi.org/10.1093/bioinformatics/btw228>.
  24. Gottlieb A, Stein GY, Ruppel E. et al. Predict: a method for inferring novel drug indications with application to personalized medicine. *Mol Syst Biol* 2011;**7**:496. <https://doi.org/10.1038/msb.2011.26>.
  25. Tang X, Zhou C, Changcheng L. et al. Enhancing drug repositioning through local interactive learning with bilinear attention networks. *IEEE J Biomed Health Inform* 2023;1–12. <https://doi.org/10.1109/JBHI.2023.3335275>.
  26. Tan Z, Chen Z, Feng S. et al. KRACL: contrastive learning with graph context modeling for sparse knowledge graph completion. In: *Proceedings of the ACM Web Conference (WWW)*, pp. 2548–59. Austin, TX, USA: ACM, 2023.
  27. Yang B, Yih W-t, He X. et al. Embedding entities and relations for learning and inference in knowledge bases. In: *3rd International Conference on Learning Representations (ICLR)*. San Diego, CA, USA: OpenReview, 2015.
  28. Sun Z, Deng Z-H, Nie J-Y. et al. RotatE: knowledge graph embedding by relational rotation in complex space. In: *7th International Conference on Learning Representations (ICLR)*. New Orleans, LA, USA: OpenReview, 2019.
  29. Chen T, Kornblith S, Norouzi M. et al. A simple framework for contrastive learning of visual representations. In: *International Conference on Machine Learning (ICML)*, pp. 1597–607. Virtual Event: PMLR, 2020.
  30. Malkiel I, Wolf L. MTAdam: automatic balancing of multiple training loss terms. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 10713–29. Punta Cana, Dominican Republic: Association for Computational Linguistics, 2021.
  31. Liu H, Zhang W, Song Y. et al. HNet-DNN: inferring new drug-disease associations with deep neural network based on heterogeneous network features. *J Chem Inf Model* 2020;**60**:2367–76. <https://doi.org/10.1021/acs.jcim.9b01008>.
  32. Zhao B-W, Lun H, You Z-H. et al. HINGRL: predicting drug-disease associations with graph representation learning on heterogeneous information networks. *Brief Bioinform* 2022;**23**:bbab515. <https://doi.org/10.1093/bib/bbab515>.
  33. Zhang M-L, Zhao B-W, Xiao-Rui S. et al. RLFDDA: a meta-path based graph representation learning model for drug-disease association prediction. *BMC Bioinformatics* 2022;**23**:516. <https://doi.org/10.1186/s12859-022-05069-z>.
  34. Cai L, Changcheng L, Junlin X. et al. Drug repositioning based on the heterogeneous information fusion graph convolutional network. *Brief Bioinform* 2021;**22**:bbab319. <https://doi.org/10.1093/bib/bbab319>.
  35. Zhao B-W, Xiao-Rui S, Peng-Wei H. et al. A geometric deep learning framework for drug repositioning over heterogeneous information networks. *Brief Bioinform* 2022;**23**:bbac384. <https://doi.org/10.1093/bib/bbac384>.
  36. Meng Y, Changcheng L, Jin M. et al. A weighted bilinear neural collaborative filtering approach for drug repositioning. *Brief Bioinform* 2022;**23**:bbab581. <https://doi.org/10.1093/bib/bbab581>.
  37. Jia X, Sun X, Wang K. et al. DRGCL: drug repositioning via semantic-enriched graph contrastive learning. *IEEE J Biomed Health Inform* 2024;PP. 1–12. <https://doi.org/10.1109/JBHI.2024.3372527>.
  38. Hu Z, Yu Q, Gao YX. et al. Drug synergistic combinations predictions via large-scale pre-training and graph structure learning. In: *27th Annual International Conference of Research in Computational Molecular Biology (RECOMB)*, pp. 265. Istanbul, Turkey: Springer Nature, 2023.
  39. Lian H, Ding P, Chao Y. et al. A weighted integration method based on graph representation learning for drug repositioning. *Appl Soft Comput* 2024;**161**:111763. <https://doi.org/10.1016/j.asoc.2024.111763>.
  40. Korczyn AD, Grinberg LT. Is Alzheimer disease a disease? *Nat Rev Neurol* 2024;**20**:245–51. <https://doi.org/10.1038/s41582-024-00940-4>.
  41. Ballard C, Aarsland D, Cummings J. et al. Drug repositioning and repurposing for Alzheimer disease. *Nat Rev Neurol* 2020;**16**: 661–73. <https://doi.org/10.1038/s41582-020-0397-4>.
  42. Angelopoulou E, Paudel YN, Papageorgiou SG. et al. Environmental impact on the epigenetic mechanisms underlying Parkinson's disease pathogenesis: a narrative review. *Brain Sci* 2022;**12**:175. <https://doi.org/10.3390/brainsci12020175>.
  43. Fletcher EJR, Kaminski T, Williams G. et al. Drug repurposing strategies of relevance for Parkinson's disease. *Pharmacol Res Perspect* 2021;**9**:e00841. <https://doi.org/10.1002/prp2.841>.
  44. Meng Y, Wang Y, Junlin X. et al. Drug repositioning based on weighted local information augmented graph neural network. *Brief Bioinform* 2024;**25**:bbad431. <https://doi.org/10.1093/bib/bbad431>.
  45. Luo H, Li M, Wang S. et al. Computational drug repositioning using low-rank matrix approximation and randomized algorithms. *Bioinformatics* 2018;**34**:1904–12. <https://doi.org/10.1093/bioinformatics/bty013>.
  46. Tang X, Cai L, Meng Y. et al. Indicator regularized non-negative matrix factorization method-based drug repurposing for COVID-19. *Front Immunol* 2021;**11**:603615. <https://doi.org/10.3389/fimmu.2020.603615>.
  47. Xiaorui S, Lun H, You Z. et al. A deep learning method for repurposing antiviral drugs against new viruses via multi-view nonnegative matrix factorization and its application to SARS-CoV-2. *Brief Bioinform* 2022;**23**:bbab526. <https://doi.org/10.1093/bib/bbab526>.
  48. Huang K, Chandak P, Wang Q. et al. A foundation model for clinician-centered drug repurposing. *Nat Med* 2024;1–13. <https://doi.org/10.1038/s41591-024-03233-x>.